Pergamon

Insect Biochemistry and Molecular Biology 33 (2003) 717–732

Insect Biochemistry and Molecular Biology

www.elsevier.com/locate/ibmb

Exploring the salivary gland transcriptome and proteome of the *Anopheles stephensi* mosquito

Jesus G. Valenzuela [a], Ivo M.B. Francischetti [b], Van My Pham [a], Mark K. Garfield [b], José M.C. Ribeiro [a,*]

[a] *National Institute of Allergy and Infectious Diseases, Medical Entomology Section, Laboratory of Parasitic Diseases, NIAID, Building 4, Room 126, 4 Center Drive, MSC 0425, NIH, Bethesda MD 20892-0425, USA*

[b] *National Institute of Allergy and Infectious Diseases, Research Technologies Branch, National Institutes of Health, Bethesda, MD 20892, USA*

Received 23 January 2003; received in revised form 26 March 2003; accepted 31 March 2003

Abstract

Anopheles stephensi is the main urban mosquito vector of malaria in the Indian subcontinent, and belongs to the same subgenus as *Anopheles gambiae*, the main malaria vector in Africa. Recently the genome and proteome sets of *An. gambiae* have been described, as well as several protein sequences expressed in its salivary glands, some of which had their expression confirmed by amino terminal sequencing. In this paper, we randomly sequenced a full-length cDNA library of *An. stephensi* and performed Edman degradation of polyvinylidene difluoride (PVDF)-transferred protein bands from salivary homogenates. Twelve of 13 proteins found by aminoterminal degradation were found among the cDNA clusters of the library. Thirty-three full-length novel cDNA sequences are reported, including a novel secreted galectin; the homologue of anophelin, a thrombin inhibitor; a novel trypsin/chymotrypsin inhibitor; an apyrase; a lipase; and several new members of the D7 protein family. Most of the novel proteins have no known function. Comparison of the putatively secreted and putatively housekeeping proteins of *An. stephensi* with *An. gambiae* proteins indicated that the salivary gland proteins are at a faster evolutionary pace. The possible role of these proteins in blood and sugar feeding by the mosquito is discussed. The electronic tables and supplemental material are available at http://www.ncbi.nlm.nih.gov/projects/Mosquito/A_stephensi_sialome/
Published by Elsevier Science Ltd.

Keywords: Salivary glands; Proteome; Electrophoresis; Hematophagy; Sialome

1. Introduction

Anopheles stephensi is the main urban malaria vector in India (Hati, 1997), belonging to the same (Celia) subgenus as the most efficient African vector, *Anopheles gambiae. An. stephensi* is susceptible both to human and to rodent malaria species such as *Plasmodium berghei*; for this reason, it is used widely as a laboratory model of *Plasmodium* development in its vector. Recently, transformation of both mosquito species has been accomplished, allowing experimental manipulation of gene expression in these mosquitoes (Catteruccia et al., 2000; Grossman et al., 2001; Nolan et al., 2002), as has been done with *Drosophila* for many years (Ashburner et al., 1998). Additionally, the genome of *An. gambiae* is nearly fully sequenced (Holt et al., 2002), providing an unprecedented opportunity to study and compare the evolution of blood feeding in the Diptera *Anopheles* and *Drosophila* (Zdobnov et al., 2002). Evolution to blood feeding involves, among other issues, the ability to obtain blood from a vertebrate host, an adaptation provided in part by the evolution of several salivary proteins that prevent vertebrate haemostasis, such as inhibitors of the clotting cascade, platelet aggregation, and vasodilators or inhibitors of vasoconstricting substances (Ribeiro, 1995). While identification of unique (e.g. not found in *Drosophila*) salivary proteins from *Anopheles* provides a clue to the evolution of blood feeding on a coarse scale, comparison of *An. gambiae* and *An. stephensi* salivary products may provide clues on a finer scale.

[*] Corresponding author. Tel.: +1-301-496-3066; fax: +1-301-402-4941.

E-mail address: jribeiro@nih.gov (J.M.C. Ribeiro).

0965-1748/03/$ - see front matter Published by Elsevier Science Ltd.
doi:10.1016/S0965-1748(03)00067-5

We have recently initiated a program outlining the sialome (set of message + proteins) of blood-sucking insects and ticks (Francischetti et al., 2002c; Valenzuela et al., 2002b; Valenzuela et al., 2002c). While these projects are descriptive in nature, they also generate hypotheses on the evolution of blood feeding in general and in the discovery of novel anti-hemostatic substances. Here we describe the sialome of *An. stephensi* and compare it with those of other mosquitoes and sand flies. Full-length cDNA information is presented for 33 novel salivary proteins of this insect. Results indicate that salivary proteins are under intense selection and can be used as robust markers for closely related species. Roles are proposed for some of the transcripts in blood feeding.

2. Materials and methods

2.1. Mosquitoes

Adult female *An. stephensi*, NIH strain, were dissected to remove the salivary glands, which were then used to make a PCR-based cDNA library using the Micro-Fast-Track mRNA isolation kit (Invitrogen, Carlsbad, CA) and the SMART™ cDNA library construction kit (BD Biosciences, Clontech, Palo Alto, CA) exactly as described (Francischetti et al., 2002c). Eighty pairs of salivary glands were used for the library.

2.2. SDS-PAGE

Sodium dodecyl sulfate polyacrylamide electrophoresis (SDS-PAGE) of 20 pairs of homogenized salivary glands of *An. stephensi* adult females was done using 1-mm thick NU-PAGE 4 to 12% gels (Invitrogen). Gels were run with MES buffer according to the manufacturer's instructions. The membrane was stained with Coomassie blue in the absence of acetic acid. Stained bands (including a negative stained band) were cut from the PVDF membrane and subjected to Edman degradation using a Procise sequencer (Perkin-Elmer, Foster City, CA). More details can be obtained in a previous publication (Francischetti et al., 2002c). To find the cDNA sequences corresponding to the amino acid sequence—obtained by Edman degradation of the proteins transferred to PVDF membranes from PAGE gels—we wrote a search program (in Visual Basic) that checked these amino acid sequences against the three possible protein translations of each cDNA sequence obtained in the mass sequencing project. For details, see Valenzuela et al., 2002c.

2.3. cDNA sequence clustering

Other procedures were as in Francischetti et al. (2002c) and in Valenzuela et al. (2002c), except that clustering of the cDNA sequences was accomplished using the CAP program (Huang, 1992). Accession numbers for the National Center for Biology Information (NCBI) databases are given, as recommended by NCBI, as gi|XXXX, where XXXX is the accession number. Accession numbers for sequences originating from the *An. gambiae* proteome (Holt et al., 2002) are given as agCP### or ebi###, where #### corresponds to the referenced gene product. BLAST searches were done locally from executables obtained at the NCBI FTP site (ftp://ftp.ncbi.nih.gov/blast/executables/) (Altschul et al., 1997). The electronic version of the complete tables (Microsoft Excel format) with hyperlinks to web-based databases and to BLAST results are available at http://www.ncbi.nlm.nih.gov/projects/Mosquito/A_stephensi_sialome/.

3. Results and discussion

3.1. Organisation of transcriptome information

To obtain insight on the salivary transcriptome of *An. stephensi* adult female mosquitoes, we randomly sequenced 1127 cDNA inserts from a salivary gland cDNA library from this insect and organised these into 362 clusters of related sequences assembled by the CAP program (Huang, 1992). Using the BLAST package of programs (Altschul et al., 1997), we compared sequences for each cluster in the database with the non-redundant protein and nucleotide sets of the NCBI and Gene Ontology databases (Ashburner et al., 2000; Hvidsten et al., 2001; Lewis et al., 2000). Translated sequences were also screened with RPSBlast for protein motifs of the combined set of Pfam (Bateman et al., 2000) and SMART (Schultz et al., 2000) databases (also known as the Conserved Domains Database (CDD)). The sequences were also compared with the proteome set of the closely related mosquito *An. gambiae* (available for FTP download at NCBI and other sites). O-glycosylation sites on the predicted proteins were obtained with the program NetOGlyc (http://www.cbs.dtu.dk/services/NetOGlyc/) (Hansen et al., 1998). Finally, we submitted all translated sequences (starting with a Met) to the Signal P server (Nielsen et al., 1997) to detect signal peptides indicative of secretion. With this information, the clustered database was annotated and classified into three categories of clusters: S, those associated with possibly secreted products; H, those possibly associated with housekeeping functions; and U, those of unknown function. Accordingly, 164 cDNA clusters containing a total of 258 sequences (23% of the transcriptome) were classified as being associated with products of category H (Fig. 1). These clusters have an average of 1.58 sequences per cluster. This contrasts with the 74 clusters containing 726 sequences (64% of the transcriptome)

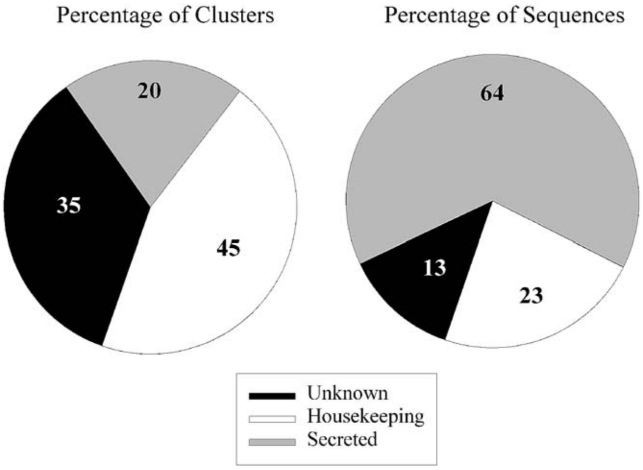


Fig. 1. The percentage of cDNA clusters (left) or sequences (right) derived from 1125 random sequences from a salivary gland cDNA library from the mosquito *An. stephensi* and classified as probably associated with either housekeeping, secretory, or unknown functions. The numbers indicate the percentage of each class of cluster or sequence.

classified as category S and providing an average of 9.93 sequences per cluster. These results of average cluster size are very different from one another ($P < 0.01$, X^2 test), and were observed also in other transcriptome analysis of salivary glands in *Ae. aegypti*, *An. gambiae* and *Ixodes scapularis* (Francischetti et al., 2002c; Valenzuela et al., 2002b; Valenzuela et al., 2002c). Finally, 141 sequences (12.5% of the transcriptome) in 123 clusters were classified as being in category U.

3.2. *Preliminary characterisation of the salivary proteome of An. stephensi*

In parallel to the organisation of the salivary gland transcriptome of *An. stephensi*, we sought to obtain information on the most abundant proteins in the salivary glands of *An. stephensi*. For this purpose, salivary gland homogenates of 20 pairs of glands were separated by SDS-PAGE, the protein transferred to a PVDF membrane, and the stained bands submitted to Edman degradation. Thirteen bands yielded useful information, 12 of which could be assigned to predicted sequences from our cluster database (Fig. 2). Other bands did not yield information, either because they were blocked at their aminoterminal, or because of the low signal. The 12 clusters associated with the observed Edman degradation results had between two and 55 sequences, with an average of 20.1 sequences per cluster—twice the average of the clusters in the S group. This result indicates that the quantity of proteins correlated with the message abundance.

The aminoterminal sequences, and the proteins they might represent, are described here in relationship to the description of the transcriptome.

3.3. *Description of S category clusters*

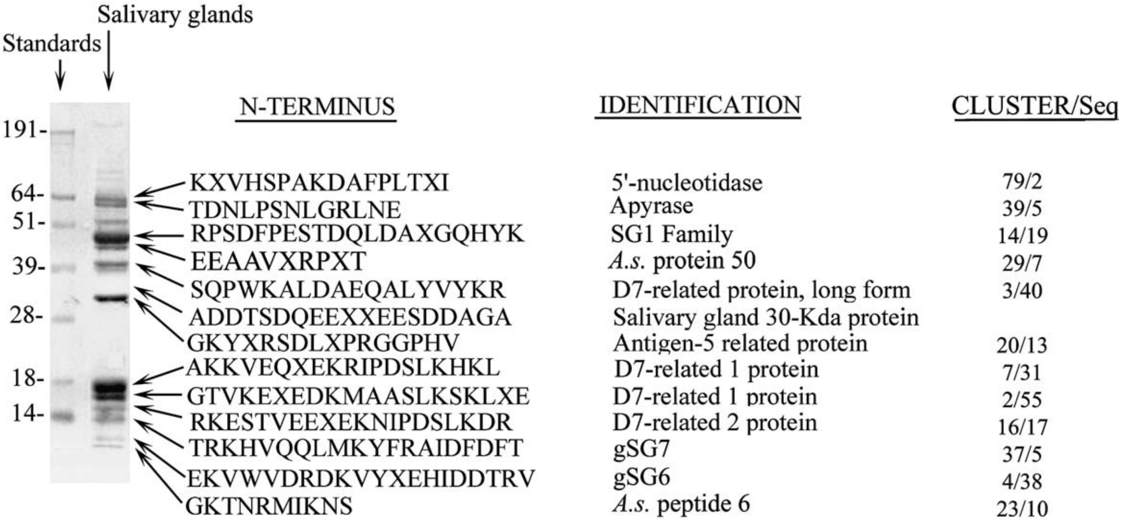
Table 1 summarises the 73 S category clusters of sequences. These clusters represent from 1 to 83 cDNA sequences each and belong to well-known families of proteins, although some do not have known function.

3.3.1. *Mucins*

The most abundant cluster codes for a protein having a Pfam domain of syndecan and a signal peptide indicative of secretion. Four other clusters were also tentatively assigned to be coding for mucin-like proteins (Table 1). Syndecans are threonine- and serine-rich transmembrane heparin sulfate proteoglycans implicated in the binding of extracellular matrix components and growth factors. The full-length sequence of a clone from this cluster is described in Table 2 as As—SG3. It codes for a protein 61% identical to *An. gambiae* SG3 protein and has repeats of TTTEEA in its carboxyterminal region responsible for the similarity to the syndecan Pfam domain, which displays triplets of Thr followed by negatively charged amino acids. Forty-six As—SG3 Thr and Ser residues are predicted to be O-glycosylated (Hansen et al., 1998). Similarly, cluster 65 (Table 1), with two sequences, may represent truncated versions of As—SG3, as it produces similarity matches by blastx to the carboxyterminal region of *An. gambiae* SG3 protein. Alternatively, it may represent an additional mucin similar to SG3. An additional abundant cluster (cluster 18, with 16 clones) codes for another protein similar to mucins. The full-length sequence yields the predicted protein named As—hyp13.5 (Table 2), with 55 predicted sites of O-glycosylation on both Ser and Thr residues. Two additional clusters may code for mucins—the singletons from clusters 140 and 293 (Table 1). Cluster 140 codes for a Thr-rich putative protein, and cluster 293 codes for the putative mucin As—hip13.4 (Table 2), which has 16 predicted O-glycosylation sites. These molecules could function as salivary mucins, helping to lubricate the salivary canal, but may also have additional activities, such as modulation of macrophages, as is the case with surface mucins of *Trypanosoma cruzi* (Acosta-Serrano et al., 2001; Ropert et al., 2002).

3.3.2. *D7 proteins*

D7 proteins, first described in *Ae. aegypti* salivary glands (James et al., 1991), are a unique family of proteins distantly related to odorant-binding proteins (Hekmat-Scafe et al., 2000) and found in the salivary glands of sand flies and mosquitoes (Valenzuela et al., 2002a). Two types of D7 proteins have been described: long D7 (28–30 kDa) found in both mosquitoes and sand flies, and short D7 (15–20 kDa) found so far only in mosquitoes (Arca et al., 1999; Valenzuela et al., 2002a). Seven members from the D7 family of proteins are reported on the NCBI database for *An. stephensi*, includ-



N-TERMINUS	IDENTIFICATION	CLUSTER/Seq
KXVHSPAKDAFPLTXI	5'-nucleotidase	79/2
TDNLPSNLGRLNE	Apyrase	39/5
RPSDFPESTDQLDAXGQHYK	SG1 Family	14/19
EEAAVXRPXT	*A.s.* protein 50	29/7
SQPWKALDAEQALYVVYKR	D7-related protein, long form	3/40
ADDTSDQEEXXEESDDAGA	Salivary gland 30-Kda protein	
GKYXRSDLXPRGGPHV	Antigen-5 related protein	20/13
AKKVEQXEKRIPDSLKHKL	D7-related 1 protein	7/31
GTVKEXEDKMAASLKSKLXE	D7-related 1 protein	2/55
RKESTVEEXEKNIPDSLKDR	D7-related 2 protein	16/17
TRKHVQQLMKYFRAIDFDFT	gSG7	37/5
EKVWVDRDKVYXEHIDDTRV	gSG6	4/38
GKTNRMIKNS	*A.s.* peptide 6	23/10

Fig. 2. Coomassie blue-stained PVDF membrane resulting from transfer of SDS-PAGE of 20 pairs of homogenized glands of adult female *An. stephensi*. Molecular weight markers are shown on the left side, as are the amino acid sequences obtained by Edman degradation for each band. The right side of the figure indicates the cluster of cDNA sequences where a match was found and the number of sequences. The band labeled ADDTS… corresponds to a negatively stained band that does not show well in the figure.

ing hamadarin, recently characterised as an inhibitor of Factor XII (Isawa et al., 2002). Table 1 indicates 10 clusters coding for proteins of the D7 family (Table 1). Table 2 describes three additional putative members of the D7 proteins, all belonging to the short form. As—D7B has 71% identity to the previously reported long-form D7clu2 salivary protein (gi|16225974) and 31% identity to the long D7 protein of *Ae. aegypti* (gi|159559) (James et al., 1991); however, a clear series of stop codons and polyA tail are found in the transcript, producing a protein of mature molecular weight similar to the short D7 proteins. As—shD7-4 (Table 2) is 84% identical to *An. arabiensis* D7r2 salivary protein, while As—shD7-5 (Table 2) is almost identical to the previously named short-form D7 salivary protein of *An. stephensi* except that it has an insert of 19 amino acid residues in the aminoterminal region. These D7 proteins may play a role in inhibiting activation of host plasma, as appears to be the case with hamadarin, or they may have different functions, as in the case with *Rhodnius* salivary lipocalins, which evolved different functions (Andersen et al., 2003; Francischetti et al., 2002a), reflecting most probably a typical scenario of gene duplication and divergence of function (Sankoff, 2001).

Consistent with the abundance of cDNA sequences in the D7 family of clusters, four bands from SDS-PAGE experiments separating *An. stephensi* salivary gland homogenate proteins produced aminoterminal sequences matching D7 proteins (Fig. 2). The sequence SPQ…, found in the 40-kDa region of the gel, matches the previously reported D7clu2, as well as the related As—D7b reported in Table 2. Because this sequence was located in a region of the gel corresponding to 40-kDa proteins, we assume it originated from the D7Clu2 protein, which has a predicted molecular weight of 35 kDa. The three remaining aminoterminal sequences match putative proteins of the short D7 family, as follows: sequence AKK… matches predicted sequences from the abundant cluster 7, which codes for the previously reported D7clu5 protein; similarly, sequence GTV… matches cluster 2, coding for the known short D7 protein from *An. stephensi* named D7r1. Sequence RKE… matches predicted sequences from cluster 16, from which a full-length clone coding for a novel short D7 protein is described in Table 2 under the name As—shD7-4. These three aminoterminal sequences are in the region of the gel expected for proteins of 15–17 kDa, in accordance with the predicted molecular weight of these short D7 proteins.

3.3.3. Antigen 5-related proteins

Closely related proteins from this family have been reported in the salivary glands of Hymenoptera sand flies (Charlab et al., 1999), tsetse (Li et al., 2001), and mosquitoes (Francischetti et al., 2002c; Valenzuela et al., 2002c). They belong to a ubiquitous family of extracellular proteins with mostly unknown function (Schreiber et al., 1997). Two cDNA clusters indicate similarities to this protein family (Table 1), both being >65% identical to a putative protein of *An. gambiae*. A full-length clone of cluster 20 (As—AG5; Table 2) is 83% identical to *An. gambiae* agCP6145 and 78% identical to the salivary antigen 5-related 1 protein of the same mosquito. The aminoterminal GKYXRSDLXPR GGXHV (The X values corresponding to predicted Cys residues) was found by Edman degradation of a PVDF-transferred gel band. This sequence is within one residue of the predicted cleavage site of the signal peptide. The

Table 1
Anopheles stephensi salivary gland cDNA clusters probably associated with secreted proteins

Cl no.[a]	No. of seqs[b]	Match to NR protein database[c]	E val[d]	Best match to CDD[e]	E val	Comments[f]
Mucin-like proteins (threonine rich)[g]						
1	83	SG3 protein	1E-59	Syndecan	2e-009	Threonine rich, As SG3(∗)
18	16	mucin-like protein	0.043	Tryp mucin	2e-006	Similar to *Ag* mucin like protein, As hyp13.5(∗)
65	2	SG3 protein	1E-27	Herpes gl	3e-004	Similar to *Ag* SG3 protein
140	1					Low complexity, Thr rich
293	1	aqCP10145	0.0008			Similar to *Ag* agCP10145, As hyp13.5(∗)
D7 proteins						
2	55	D7-related 1 protein	2E-90			As D7r1
24	9	D7-related 1 protein	1E-78			Very similar to As D7r1, As shD7-5(∗)
3	40	long form D7clu2 salivary	1E-103			As long form D7clu2
8	27	long form D7clu2 salivary	3E-63			As long form D7clu2, As D7B(∗)
122	1	long form D7clu2 salivary	2E-21			As long form D7clu2, truncated clone?
7	31	short form D7clu5 salivary	4E-92			As short form D7clu5
16	17	short form D7r2	4E-58			Similar to *As* D7r2, As shD7-4(∗)
129	1	short form D7r2	1E-16			Similar to *As* D7r2
51	4	D7-related 3 protein	1E-43			Similar to *Ag* D7r3
113	1	Hamadarin	2E-45			As short form D7clu5, hamadarin
Antigen-5 related protein						
20	13	Antigen 5-related protein	1E-102	SCP	2e-029	Similar to *Ag* antigen 5 rel1 protein, As AG5(∗)
112	1	Antigen 5-related protein	2E-23	SCP	3e-006	Similar to *Ag* antigen 5 rel1 protein
Similar to previously described salivary *An. Gambiae* proteins of unknown function 30 kDa allergen						
42	4	30 kDa protein	5E-26	MAGE	6e-009	Similar to *Ag* 30 kDa protein, As GE(∗)
SG1 family of anopheline salivary proteins						
59	3	GSG1b protein	5E-27	7tm 1	0.007	Similar to *Ag* gSG1b, As SG1C(∗)
29	7	Hypothetical protein	8E-67			Similar to *Ag* hypothetical protein, As SG1D(∗)
86	2	Salivary gland1-like 3 protein	4E-12			Similar to *Ag* sg 1 like 3 protein
35	6	SG1 protein	3E-29			Similar to *Ag* SG1
41	5	SG1 protein	1E-57			Similar to *Ag* SG1 protein – As SG1B(∗)
91	2					Unknown – As SG1A(∗)
14	19	Salivary gland1-like 3 protein	4E-76			Similar to *Ag* SG1-like 3 protein
40	5	TRIO protein	2E-42			Similar to *Ag* TRIO protein, As TRIO(∗)
94	1	TRIO protein	1E-34			Similar to *Ag* TRIO prtn

(*continued on next page*)

Table 1 (*continued*)

Cl no.[a]	No. of seqs[b]	Match to NR protein database[c]	E val[d]	Best match to CDD[e]	E val	Comments[f]
SG2 family of anopheline salivary proteins						
12	23	SG2 protein – agCP6138	2E-31			Similar to *Ag* SG1 protein, As SG2B(*)
25	8	gSG2 protein – agCP6166	1E-30	STT3	2e-004	Similar to *Ag* gSG2-like protein, As SG2A(*)
31	6	gSG2 - agCP6166	1E-29			Similar to *Ag* gSG2-like protein
36	5	gSG2 - agCP6166	9E-30			Similar to *Ag* gSG2-like protein
61	2	SG2 protein - agCP6138	9E-31			Similar to *Ag* SG2 protein
93	1	SG2 protein - agCP6138	2E-31			Similar to *Ag* SG2 protein
97	1	SG2 protein - agCP6138	1E-32			Similar to *Ag* SG2 protein
109	1	SG2 protein - agCP6138	3E-08			Similar to *Ag* SG2 protein
Similar to reported salivary hypothetical proteins from *An. gambiae* (unique to anophelines)						
33	6	gSG5 protein	5E-43			Similar to *Ag* gSG5 protein
4	38	gSG6 protein	3E-42			Similar to *Ag* gSG6, As gSG6(*)
37	5	gSG7 protein	2E-49			Similar to *Ag* gSG7
15	18	gSG7 protein	5E-53	DiHfolate red	0.006	Similar to *Ag* gSG7 protein, As gSG7(*)
34	6	gSG8 protein	8E-55	Borrelia orfA	1e-005	Similar to *Ag* gSG8 protein
32	6	hypothetical protein	1E-08			Similar to *Ag* hypothetical protein, As hyp8.2(*)
53	3	hypothetical protein 10	1E-12			Similar to *Ag* hypothetical protein 10, As hyp10
26	8	hypothetical protein 12	5E-08			Similar to *Ag* hypothetical protein 12, As hyp12
30	6	hypothetical protein 15	2E-12			Similar to *Ag* hypothetical protein 15, As hyp15
23	10	hypothetical protein 15	3E-12			Similar to *Ag* hypothetical protein 15
Similar to *An. gambiae* proteins not previously described in salivary glands						
11	24	agCP2783	4E-55			Similar to *Ag* agCP2783, As hyp16(*)
17	16	agCP12351	8E-49			Similar to *Ag* agCP12351, As hyp37.3(*)
128	1	agCP4078	7E-50			Similar to *Ag* agCP4078, As sal lipase(*)
210	1	agCP13847	3E-84			Similar to *Ag* agCP13847
272	1	agCP5790	1E-37			Similar to *Ag* agCP5790, As hyp6.3(*)

aminoterminal sequence of As—AG5 was located in the 30-kDa region of the gel, in accordance with the 28-kDa prediction of the molecular weight of the mature protein.

3.3.4. 30-kDa antigen

This protein was first described in *Aedes* mosquitoes (Simons and Peng, 2001) but is also found in *An. gambiae* (Francischetti et al., 2002c). It has a long region of low amino acid complexity, consisting mainly of Gly and Glu residues. The function of this protein is unknown.

The sequence ADDTS… was found in a negative staining band in the gel shown in Fig. 2, in a region corresponding to 34-kDa retention. This sequence does

Table 1 (*continued*)

Cl no.[a]	No. of seqs[b]	Match to NR protein database[c]	E val[d]	Best match to CDD[e]	E val	Comments[f]
Enzymes probably linked to anti-hemostatic activities						
21	11	peroxidase	0	An peroxidase	2e-045	Salivary peroxidase
277	1	Peroxidase	3E-55	An peroxidase	6e-031	Peroxidase, truncated clone
39	5	apyrase	1E-79	5 nucleotidase	3e-056	Similar to *Ag* apyrase
79	2	putative 5′-nucleotidase	2E-88	5—nucleotidase	1e-054	Similar to *Ag* putative 5′-nucleotidase, As—sal—apyrase (∗)
137	1	apyrase	1E-100	5 nucleotidaseC	2e-028	Similar to *Ag* apyrase, truncated clone
252	1	putative 5′-nucleotidase	2E-80	5 nucleotidaseC	3e-040	Similar to *Ag* putative 5′nucleotidase, truncated
Serine protease inhibitors						
5	38	anophelin	2E-15	ion trans	6e-006	Anophelin, As Anophelin (∗)
10	24	anophelin	2E-15	Transmembrane4	0.006	Anophelin
110	1					Anophelin (low blast match)
95	1	Chymotrypsin inhibitor	3E-10	TIL	2e-015	Similar to *Apis* chymotrypsin inhibitor, As—protease—inhib(∗)
Enzymes linked to sugar meals						
6	36	Maltase precursor	0	Aamy	3e-074	Maltase
101	1	Possible maltase L precursor	3E-11	Aamy	2e-011	Maltase, second enzyme
Possibly related to immunity						
9	27	Putative 56.5kD protein	1E-44	RTX	8e-004	Similar to putative 56.5kDa secreted protein, As—hyp53.7 (∗)
19	14	Lysozyme precursor	2E-68	LYZ1	1e-049	Lysozyme, As lysozyme (∗)
327	1	gram negative binding protein	3E-39			Gram negative binding protein
330	1	Lectin	4E-38	GLECT	5e-016	Galectin, As hyp21.2(∗)
Unknown function or family						
139	1					Unknown
171	1	expressed sequence AW54	0.09			Unknown
284	1					Unknown, As hyp11.9(∗)
318	1					Unknown, As hyp13.3(∗)
343	1					Unknown
357	1					Unknown

(∗) Indicates full length sequence obtained and summarized in Table 2.

[a] cDNA clones were clustered by the program CAP assembler (Huang, 1992).

[b] Number of sequenced clones in cluster.

[c] Best protein match by blastX to the non redundant protein database of NCBI.

[d] Significance of the match.

[e] Best match by RPSblast to the Conserved Domains database.

[f] As and *Ag* refer to *An. stephensi* and *An. gambiae*.

[g] Indicates association of putative translation products with function or families of proteins.

not produce a clear match to any of the translation products of our database, but closely matches the salivary 30-kDa protein of *An. gambiae* and may, accordingly, represent the homologous protein in *An. stephensi*.

3.3.5. *The SG1 family*

This family of anopheline salivary proteins, described as SG1 or gSG1 proteins (Arca et al., 1999; Lanfrancotti et al., 2002), does not yield significant similarities (by blastp) to other proteins in the NCBI database except

Table 2
Characterization of 33 full length cDNA clones from a salivary gland library of *An. stephensi*

	SP[a]	MW[b]	Pl[c]	MW[d]	pl[e]	Best match to NR protein database[f]	E val[g]	Comment
Mucin-like proteins								
As—SG3	18	20.9	4.37	19.1	3.31	SG3 protein	1E-59	Similar to *Ag.* SG3 protein
As—hyp13.5	23	16.1	4.43	13.5	4.18	mucin-like protein	1E-42	Mucin like protein
As—hyp13.4	26	16.3	3.58	13.4	3.22	mucin	2E-12	Putative mucin
D7 family								
As—D7B	18	18.5	7.7	16.5	8.26	D7 protein	9E-64	Short D7 similar to long form
As—shD7-4	21	18.4	5.25	16.13	4.95	short form D7r2	3e-076	Short D7 protein
As—shD7-5	24	17.3	8.77	14.68	8.69	D7-related 1 protein	1e-062	Short D7 protein – aminoterminal insert
Antigen-5 related protein								
As—AG5	21	29	9.05	26.8	9.17	antigen 5-related 1 pro	1E125	Antigen 5 related protein – aminoterminal
Similar to 30 kDa allergen								
As—GE	19	28.5	4	26.36	3.95	30 kDa protein	2E-32	Similar to *Aedes* 30 kDa allergen
SG1 Family (unique to anophelines)								
As—TRIO	25	43.8	7.01	40.95	7.11	TRIO protein	5E-74	Similar to *Ag* TRIO protein
As—SG1C	22	44.3	6.73	42	7.26	gSG1b protein	1E-119	Similar to *Ag* gSG1b protein
As—SG1D	24	46.9	9.38	44.27	9.49	hypothetical protein	6E-74	Similar to *Ag* gi|18873404 hypothetical protein – amino terminal sequence EEAAVXRPXT found in SDS-PAGE
As—SG1B	22	48.2	6.24	45.7	6.06	SG1 protein	1E-97	Similar to *Ag* SG1 protein
As—SG1A	26	50.2	9.4	47.5	9.43	SG1 protein	1I-76	Similar to *Ag* SG1 protein
SG2 Family (unique to anophelines)								
As—SG2B	20	11.7	5.28	9.7	4.41	agCP6138	2E-09	Similar to *Ag* SG1 protein
As—SG2A	18	13.2	11	11.3	10.8	agCP6166	7E-17	Similar to *Ag* gSGw-like protein
Similar to reported salivary hypothetical proteins from *An. gambiae* (unique to anophelines)								
As—hyp15	29	8.2	10.04	5	11.7	hypothetical protein 15	6E-13	Similar to *Ag* hypothetical protein 15 – amino terminal GKTNRMIKNS found in SDS-PAGE
As—hyp10	34	11.1	5.58	7.5	4.92	hypothetical protein 10	6E-13	Similar to *Ag* hypothetical protein 10
As—hyp8.2	18	10.1	4.65	8.2	4.51	hypothetical protein	3E-09	Similar to *Ag* CAA76819 – hypothetical protein
As—gSG6	28	12.9	5.1	10.01	5.31	gSG6 protein	2e-047	Putative 10 kDa protein amino terminalEKVWVDRDKVYXEHIDXTRV found in SDS-PAGE
As—gSG7	26	16.6	9.62	13.8	9.71	gSG7 protein	1E-52	Similar to *Ag* gSG7
Similar to *An. gambiae* proteins not previously described in salivary glands								
As—hyp37.3	16	39.2	9.51	37.34	9.44	agCP12351	1E-55	Similar to *Ag* agCP12351 – hypothetical 37.3
As—hyp16	25	18.5	5.6	16	5.59	agCP2783	4E-55	Similar to *Ag* agCP2783 – hypothetical protein
As—hyp6.3	17	8.2	6.54	6.3	8.1	agCP5790	8E-35	Very similar to *Ag* agCP5790
Salivary lipase								
As—sal—lipase	28	35.8	5.36	32.8	5.24	pancreatic lipase	1E-151	Lipase
Enzymes linked to anti-hemostatic activities								
As—sal—apyrase1	27	64.3	6.77	61.3	6.5	putative 5′-nucleotidase	0	Similar to *Ag* putative 5′-nucleotidase – aminoterminal KXVHSPAKDAFPLTXI found in SDS-PAGE

(*continued on next page*)

Table 2 (*continued*)

	SP[a]	MW[b]	Pl[c]	MW[d]	pl[e]	Best match to NR protein database[f]	E val[g]	Comment
Serine protease inhibitors								
As—anophelin	21	11	4.12	8.9	4.04	Anophelin	5E-16	Anophelin – anti-thrombin
As—protease—inhib	22	9.4	8.06	7.1	7.63	venom protein	7E-11	Trypsin/chymotrypsin inhibitor
Possibly related to immunity								
As—lysozyme	22	15.4	8.8	13.21	8.62	Lysozyme precursor	2e-069	Lysozyme
As—hyp53.7	22	56.3	5.79	53.7	5.25	putative 56.5 kD	3E-71	Similar to *Aedes* putative 56.5 kDa secreted
As—hyp21.2	22	23.7	9.47	21.2	9.48	lectin	2E-99	Secreted galectin
Putative proteins with low similarity to other known proteins								
As—hyp13.3	19	15.3	5.28	13.3	5			Putative 13.3 kDa protein
As—hyp11.9	21	14.1	4.33	11.9	4.33			Putative 11.9 kDa protein

[a] Last amino acid position in signal peptide predicted by the SignalP program (Nielsen et al., 1997).

[b] Predicted molecular weight of the protein.

[c] Predicted isoelectric pH of the reduced protein.

[d] Predicted molecular weight of the mature protein.

[e] Predicted isoelectric pH of the reduced mature protein.

[f] Best match by BlastP to the NR protein database of the NCBI.

[g] E value of the best match.

among its own members. This family also includes, distantly, the salivary *An. gambiae* TRIO protein (Francischetti et al., 2002c). Nine cDNA clusters of the *An. stephensi* salivary gland library yielded similarities to *An. gambiae* proteins annotated as members of the SG1 family (Table 1). Clusters 40 (with five sequences) and 94 (one sequence) may represent full-length and truncated versions of a homologue of *An. gambiae* salivary TRIO protein (Francischetti et al., 2002c). Similarly, clusters 35 and 41 may represent full-length and truncated versions of the homologue of *An. gambiae* SG1 protein. Five full-length sequences were obtained from clones belonging to different clusters of this family of proteins, representing the putative proteins As—TRIO, As—SG1A, As—SG1B, As—SG1C, and As—SG1D (Table 2). All proteins have a clear signal peptide indicative of secretion and a predicted mature molecular weight between 40.9 and 47.5 kDa.

The aminoterminal sequence RPS… was found in a protein band of the SDS-PAGE experiment reported in Fig. 2, corresponding to transcripts of the abundant cluster 14, with 19 sequences coding for a protein similar to *An. gambiae* described as salivary gland1-like 3 protein. Another member of the SG1 family is represented in the experiment reported in Fig. 2, indicated by the aminoterminal sequence EEAA…, which matches cluster 29 with seven sequences, and the full-length sequence contained in As—SG1D. Both aminoterminal sequences were found in the 46–48 kDa region of the gel, in agreement with the expected molecular weights of the mature proteins.

3.3.6. The SG2 family

This family, first described in *An. gambiae* as SG2, gSG2, and SG2-like protein (Arca et al., 1999; Lanfrancotti et al., 2002) consists of proteins rich in Gly or Asn residues and having 114–168 amino acid residues. High similarity matches are only produced among these anopheline proteins; the remaining low-score matches are obtained only when the BLAST filter for low-complexity regions is removed. It is interesting that *Ixodes scapularis* also contains glycine-rich peptides of equivalent size (Valenzuela et al., 2002b). Their function is also unknown.

The *An. stephensi* transcriptome produced eight clusters of cDNA sequences having similarity to this protein family, including two abundant clusters (cluster 12 and 25, with 23 and eight clones each) giving strong similarity to *An. gambiae* SG2 and gSG2 proteins (Table 1). Individual clones from these two clusters were fully sequenced; their putative translated protein properties are summarised in Table 2. Of the remaining six clusters shown in Table 1 for SG2 proteins, five are very similar to either SG2 or gSG2 (clusters 25, 31, 61, 93 and 97), yielding products with only a few amino acid changes from the predicted proteins from either cluster 12 or 25, possibly representing alleles of these genes or, alternatively, sequencing error. Their inclusion in this work may stimulate those looking for polymorphism in salivary gland proteins of blood-sucking insects (Lanzaro et al., 1999). Cluster 109, however, is similar to the SG2 protein but lacks predicted amino acids from positions 23–101 and may represent an alternative transcript of the *An. stephensi* gene homologue to *An. gambiae* SG2 protein.

3.3.7. Hypothetical proteins and glandins apparently unique to anophelines

Ten abundant cDNA clusters from *An. stephensi* salivary gland library (containing from 3–38 clones each) yielded similarities to putative salivary *An. gambiae* proteins previously described as glandins (Lanfrancotti et al., 2002) or salivary hypothetical proteins (Francischetti et al., 2002c). Similar to SG1 and SG2 proteins, these glandins and hypothetical proteins do not yield significant matches to other proteins in the NR protein database. Some of these clusters may represent allelic variants. Full-length sequence information for six of these putative proteins from *An. stephensi* is reported in Table 2. They all contain signal peptide indicative of secretion and code for proteins of relatively small molecular weight, varying from 5–15 kDa. Aminoterminal sequence of PVDF-transferred SDS-PAGE protein bands yielded results matching three proteins in this group, As—hyp15, As—gSG6, and translation products of cluster 37. As—hyp15 has a predicted signal peptide cleavage at position 28–29, while the found aminoterminal GKTN… is at position 33–34. As—gSG6 has a signal peptide correctly predicted, as indicated by the observed aminoterminal sequence EKV…. The function of these short proteins is unknown.

3.3.8. Additional hypothetical proteins not previously described in mosquito salivary gland transcriptomes, but similar to predicted An. gambiae proteins

The *An. stephensi* salivary gland cDNA library yielded five clusters of sequences similar to predicted proteins from the *An. gambiae* genome but not previously described in salivary gland transcriptomes. Full-length sequence information is provided for three clones from these clusters and is represented in Table 2 as As—hyp37.3, As—hyp16, and As—hyp6.3. These three putative proteins have predicted signal peptides indicative of secretion. Of interest, As—hyp6.3, with a predicted mature molecular weight of 6.3-kDa, has significant matches to the aminoterminal region of proteins annotated as dehydrogenases. The function of these three putative proteins is unknown.

An additional clone (As—sal—lipase; Table 2) from this group was fully sequenced and found to be substantially similar to proteins annotated as triacylglycerol lipases. The existence of a salivary lipase was not previously described in the saliva of blood-feeding insects, although a phospholipase C with specificity to the phospholipid platelet-activating factor (PAF) was recently described (Ribeiro and Francischetti, 2001). PAF hydrolysis by salivary gland homogenates of anophelines was not found in the same work. The presence and role of the salivary lipase of *An. stephensi* remains to be investigated.

3.3.9. Enzymes associated with anti-hemostatic activities

Salivary peroxidases were found in the salivary glands of anopheline mosquitoes, where they act as vasodilators (Ribeiro and Valenzuela, 1999; Ribeiro and Nussenzveig, 1993; Ribeiro et al., 1994). Two cDNA clusters matching *An. albimanus* salivary peroxidase were obtained, one of which is a carboxyterminus truncated clone. These two clusters, representing a total of 12 sequences, predict peptide sequences ~80% identical to *An. gambiae* protein ebiP593, which may represent the salivary peroxidase of this mosquito.

Apyrases are enzymes ubiquitously found in the salivary glands of blood-feeding insects and ticks. These enzymes, belonging to different protein families, degrade the neutrophil-inducing substance ATP and the platelet-aggregating nucleotide ADP to AMP, presumably facilitating blood feeding. *Ae. aegypti* apyrase is a member of the 5′-nucleotidase family (Champagne et al., 1995). In *An. gambiae*, two such genes are expressed in the salivary glands and annotated as apyrase and 5′-nucleotidase; however, both could actually be coding for proteins with apyrase activity (Arca et al., 1999; Lombardo et al., 2000). *An. stephensi* salivary cluster 39 codes for a protein having 79% identity with the *An. gambiae* salivary apyrase (gi|4582524), while cluster 79 codes for a protein having 80% identity to the *An. gambiae* salivary putative 5′-nucleotidase (gi|4582528). The other two clusters, containing one sequence each, represent truncated clones matching the carboxyterminal region of both above-named gene products. Accordingly, cluster 39 and 79 code for proteins homologous to *An. gambiae* salivary apyrase and putative 5′-nucleotidase, respectively. In support of the existence of two salivary apyrases in *An. stephensi* is the previous finding of two ATP and ADP, but not AMP, hydrolysing bands visualised in isoelectric focusing gel experiments (Mathews et al., 1996). Additionally, aminoterminal sequences matching both putative salivary apyrase gene products were found in the 60-kDa region of SDS-PAGE gels of *An. stephensi* (Fig. 2). The full-length clone of the *An. stephensi* homologue of *An. gambiae* gi|4582528 is reported in Table 2 as As—sal—apyrase1.

3.3.10. Antithrombin (anophelin) and anti-protease products

The *An. stephensi* salivary transcriptome has three cDNA clusters coding for proteins similar to antithrombins of the anophelin family (Valenzuela et al., 1999). Close inspection of these clusters (Table 1, clusters 5, 10 and 110) reveals that assembled clusters 5 and 10 have identical nucleotide sequences in their coding regions and may have been built as two individual clusters due to the presence of sequences having relatively high number of undetermined nucleotides in the overall sequence. Cluster 110, a singleton, is a truncated anophelin clone. Both clusters 5 and 10 give matches to an *An. gambiae* putative protein sequence and to two *An. gambiae* protein sequences, cE5 and F1 (Arca et al., 1999). *An. gambiae* sequence F1 appears to be a truncation of sequence cE5 and not a novel gene product, while sequence cE5 matches the predicted genomic sequence. It appears that *An. stephensi* has a single highly transcribed anophelin gene homologue (68 total transcripts in this *An. stephensi* library). The full-length putative *An. stephensi* anophelin peptide (As—anophelin; Table 2) is 49 and 42% identical to the homologues of *An. gambiae* and *An. albimanus*, respectively. It will be interesting to verify whether anophelin peptides from old world mosquitoes have the same high affinity to thrombin as *An. albimanus* anophelin (Francischetti et al., 1999).

An. stephensi cluster 95 (Table 1) codes for a protein similar to *Apis* chymotrypsin inhibitor and could act as an elastase inhibitor, an activity that would confer antiïnflammatory activity (Gompertz and Stockley, 2000) or inhibit blood clotting. The full-length clone (As—protease—inh; Table 2) predicts a secreted protein with 10 conserved cysteines found in some serine protease inhibitors and in the procoagulant plasma protein Von Willebrand factor. This protein has relatively weak similarity (45% identity at the amino acid level) to a predicted protein of *An. gambiae*.

3.3.11. Sugar-meal digestion

Mosquito salivary glands express enzymes such as maltases (James et al., 1989; Marinotti et al., 1990) that help sugar-meal digestion. The homologue of the maltase gene of *Ae. aegypti* (James et al., 1989) was found in the cluster database (cluster 6, with 36 sequences), as well as cluster 101, a singleton, coding both for putative proteins with the Pfam motif for alpha-glucosidase (Table 1). Cluster 101 is not a truncated clone, however, because it matches unspecified *Anopheles* and *Drosophila* proteins (~600 amino acid residues long) starting at amino acid positions 20 or 40. It is possible that *An. stephensi* expresses two gene products related to maltases in its salivary glands.

3.3.12. Putative immunity-related products

Lysozyme, an antibacterial enzyme first described as a mosquito salivary activity in *Ae. aegypti* (Rossignol and Lueders, 1986), is associated with the abundant cluster 19, with 14 sequences. Salivary lysozyme may help to deter bacterial growth in sugar meals of mosquitoes, which are stored in the crop. The full-length An—lysozyme is 84% identical to *An. gambiae* LYC—ANOGA at the amino acid level (Table 2), and has a signal peptide indicative of secretion.

The abundant cluster 9, with 27 sequences, codes for a protein with similarity to a salivary protein from *Ae. aegypti* described as a putative 56.5-kDa protein (Valenzuela et al., 2002c). The full-length sequence from a clone from this cluster (As—hyp53.7; Table 2) codes for a protein 33% identical (BLAST E value, 3E-71) to this *Aedes* putative protein (gi|18568292). No similar proteins were found for *An. gambiae* in two independent projects searching for salivary proteins in *An. gambiae* (Arca et al., 1999; Francischetti et al., 2002c; Lanfrancotti et al., 2002), and no similar proteins were found in the putative proteins from the genome of *An. gambiae*, which covers 85% of this mosquito's genome (Holt et al., 2002); however, when the protein sequence was compared to the *An. gambiae* genome (using the program tblastn), a putative protein containing 70% amino acid a identity was found in the genomic sequence gb|AAAB01008960.1| spanning as a single exon from position 7921116 to 7922651, corresponding to amino acid residues 3–516 in this *An. stephensi* putative protein with 516 amino acids. This protein has the Pfam RTX signature found in cytolytic bacterial toxins and could act as an antibacterial protein.

Two additional clusters may be involved with salivary immunity function. These are cluster 327, coding for a Gram-negative binding protein, and cluster 330, coding for a galectin. Both of these clusters code for putative secreted proteins. Expression of the message coding for the Gram-negative binding protein in mosquito salivary glands has been described before for *An. gambiae* (Dimopoulos et al., 1997) and *Ae. aegypti* (Valenzuela et al., 2002c). The presence of a galectin in any insect saliva is a novel finding, although salivary glands of anophelines are known to contain hemaglutinating compounds (Gooding, 1972; Metcalf, 1945) that may function as an adjuvant for immunity reactions (Vasta et al., 1999) or help blood feeding by concentrating the blood meal (Vaughan et al., 1991). As—galectin (Table 2) is 81% identical at the amino acid level with a protein of similar size from *An. gambiae* (agCP6926).

3.4. Putative messages associated with secretory products from the U category

Six cDNA singletons code for transcripts predicted to have a signal peptide indicative of secretion but that otherwise do not match other known proteins, even when the BLAST filter to exclude low-complexity sequences was removed. Full-length cDNA information was obtained for two of these transcripts, As—hyp13.3 and As—hyp11.9 (Table 2). Both putative proteins have a signal peptide indicative of secretion. No similarity to known proteins was found for either protein. When the protein sequences were compared to the *An. gambiae* genome (using tblastn), they produced matches indicating these are novel proteins in *An. gambiae* that have not been previously annotated. Accordingly, As—hyp13.3 produces a match of 48% identity at the amino acid level with the *An. gambiae* genome sequence gb|AAAB01008964.1| from 6017212 to 6017586 corresponding to amino acid residues 17–140 of this 140 amino acid-long protein from *An. stephensi*. Similarly, the protein sequence predicted in As—hyp11.9 produces a match with 67% identity to the translated *An. gambiae* genomic sequence gb|AAAB01029954.1| from 262 to 621, corresponding to amino acid residues 8–128 in this 128-residue predicted *An. stephensi* protein. Their function is unknown.

3.5. Description of H category clusters of transcripts

Of the 164 H category clusters in the salivary transcriptome of *An. stephensi* (Supplemental table), 63 clusters code for proteins involved in protein synthesis including ribosomal proteins, elongation factors and transcription–initiation factors. Thirty-five clusters are associated with energy metabolism, including several mitochondrial enzymes (cytochromes and ATP synthases) and enzymes from the glucolysis, glucose-6-P and Krebs cycle pathways (transaldolase, transketolase, aconitase and several dehydrogenases). Eleven clusters are associated with possible transcription factors, including imaginal tissue growth factor, transcriptional activators, cellular repressor of E1A-stimulated genes, a product similar to the *D. melanogaster* male-specific lethal which may be involved in differentiation of the female salivary gland and death-associated protein 1. Ten clusters are associated with signal transduction pathways including small GTPases, calmodulin, protein kinase C inhibitors, and G proteins; another 10 clusters are associated with membrane ATPases involved in ion transport, such as subunits of the V-ATPase and $Na^+ + K^+$ ATPase, and other membrane proteins of unknown function. Nine clusters are associated with products possibly involved in protein folding and export, such as cyclophylin, thiol reductase, thioredoxin peroxidase, protein disulphide isomerase, the V-snare protein, and a translocon protein. Six clusters are associated with cytoskeletal proteins, such as troponin, actin, tensin, dynactin and tubulin. Two clusters are associated with lipid metabolism, including an apolipophorin similar to the apolipophorin II of the tobacco hornworm *Manduca sexta* and a protein with similarity to sterol transporter and beta-oxidation proteins. Of further interest are clusters coding for products with similarity to enzymes associated with protein glycosylation, such as polypeptide N-acetylgalactosaminyltransferase and mannose-1-phos-

Table 3
Identity, at amino acid level, between housekeeping and putative secreted salivary proteins of *An. stephensi* and *An. gambiae*

Ag sequence	Description	% identity	Match position[a]
Housekeeping genes			
gi|21297563	ribosomal protein S17	100	1–133
gi|21296974	ribosomal protein L19	99	1–154
gi|21287970	ribosomal protein S26	99	1–119
gi|21292099	ribosomal protein S4e	98	1–162
gi|21296658	ribosomal protein S3a	97	1–272
gi|21294671	Actin 5C	96	24–288
gi|21297098	Ribosomal protein S25	96	17–142
gi|21297821	ribosomal protein S20	96	1–135
gi|21293011	60S ribosomal protein L	95	1–146
gi|21302101	ribosomal protein L13a	95	2–168
gi|21296028	ribosomal protein L26	94	3–171
gi|21287770	ribosomal protein L27	94	1–127
gi|21288798	ribosomal protein L35	93	1–132
gi|21296878	Thiol reductase	91	6–189
gi|21288144	ribosomal protein L21	91	1–168
gi|21293738	Homolog of Dm RE1810	88	28–217
gi|21299211	ribosomal protein L14	86	33–204
gi|21301822	ribosomal protein L12	85	1–140
gi|21292994	ribosomal protein L11	76	2–187
Average ± S.D.		93.11	5.93
Salivary gland genes			
gi|21300976	Probable mucin	92	11–120
gi|21288705	maltase	88	8–294
gi|2497784	Iyzozyme precursor	84	1–140
gi|21300145	D7-related 2	83	4–172
gi|21294898	peroxidase	79	79–214
gi|4582528	Putative 5′-nucleotidase	79	19–219
gi|13537666	gSG6 protein	78	1–114
gi|18389883	antigen 5-related 1	78	1–178
gi|21300324	gSG7	70	11–148
gi|21291106	sG7 family	69	11–149
gi|13537674	gSG8 protein	69	5–127
gi|21294389	hypothetical protein	68	7–195
gi|4582524	apyrase	68	27–246
gi|21299164	SG2 protein	65	12–122
gi|4538891	D7-related 3	65	5–109
gi|4210617	SG2 protein	64	1–114
gi|21298718	SG3 protein	61	13–207
gi|18389893	mucin-like protein	61	18–178
gi|4538887	D7-related 1 protein	60	1–164
gi|21300147	Long D7	57	8–162
gi|13537662	gSG5 protein	51	1–212
gi|18389891	Long D7	50	1–306
gi|21289292	novel protein	50	4–205
gi|21299419	gSG2-like protein	50	9–180
gi|21300288	cE5 protein	49	2–105
gi|21299419	gSG2-like protein	48	9–180
gi|21294236	SG1 family	46	118–328
gi|13537664	gSG1b protein	44	6–166
gi|21301831	30 kDa protein	43	7–172
agCP12812	TRIO protein	42	5–176
gi|18873404	Hypothetical protein	40	1–363
gi|4210615	SG1 protein	35	1–153
Average ± S.D.		62.06	15.40

[a] Indicates amino acid residues in *An. gambiae* protein matching the *An. stephensi* homologue.

phate guanylyltransferase, probably associated with glycosylation of secreted proteins. A cluster coding for a product with high similarity to alpha-endosulfine—the endogenous peptide that regulates the ATP-dependent K^+ channel—was also found. Finally, a cDNA coding for a product with similarity to enzymes annotated as gamma-glutamyl hydrolase suggests production of an enzyme that could inactivate enzymes involved with blood clotting; however, no clear signal peptide indicative of secretion was observed, and this cluster was annotated in the H category.

3.6. *Comparison of protein sequence identities between An. stephensi and An. gambiae gene products*

It has been proposed that, in American sand flies, adult female salivary gland proteins are under strong selection due to the deleterious effect that host immunity has on feeding (Lanzaro et al., 1999). It may also be that the salivary gland genes involved in blood feeding are rapidly evolving to adapt to a different repertoire of hosts, although the primary host for both *An. gambiae* and *An. stephensi* is human. To test whether salivary gland genes are evolving at a faster pace than housekeeping genes, we compared the degree of identity, at the translated amino acid level, between category H (mostly ribosomal gene products) and category S genes. For this comparison (Table 3), we used the *An. gambiae* protein data set recently submitted to NCBI and *An. stephensi* sequences that originated from two or more cDNA sequences and that gave >100 amino acid residues of match to the *An. gambiae* sequences when these were compared by blastp with the filter removed. Both the average and the variance of the two data sets were very significantly different ($P < 0.0001$). The H genes had an average of 93.11 ± 5.93% identity, while the S genes had 62.4 ± 15.4 (average ± S.D.; averages tested by *t*-test with non-equal variances; variances tested by the *F*-test). We conclude that the salivary gland genes of these two anophelines of the Celia subgenus are rapidly evolving in comparison with housekeeping genes. These results support the idea that S genes may be good markers to assess phylogeny between closely related species as has been demonstrated with closely related *Rhodnius* triatomines using the salivary hemeproteins (Soares et al., 1998; Soares et al., 2000).

4. Final remarks

One striking observation when contemplating hematophagous animal salivary gland transcriptomes is our inability to assign a role for most of the gene products. Expression of these proteins in large amounts and screening for their possible role in multiple bioassays will facilitate understanding how these organisms have adapted to disarm host hemostasis and inflammation, as was recently done for the D7 protein hamadarin (Isawa et al., 2002); Ixolaris, the tissue factor pathway inhibitor of the tick *Ixodes scapularis* (Francischetti et al., 2002b); the tick histamine-binding proteins (Paesen et al., 2000); and *Rhodnius* biogenic amine-binding protein (Andersen et al., 2003). Additionally, 18 of the proteins described in this paper appear to be unique to anopheline mosquitoes. These include two highly expressed members of the SG1 family (Fig. 2) that could be good markers of anopheline exposure, such as has been accomplished with ticks (Schwartz et al., 1990) and sand flies (Barral et al., 2000).

Acknowledgements

We are grateful Drs Robert Gwadz, Thomas Kindt, and to Dr Thomas Wellems for encouragement and support, and to Ms. Nancy Shulman for editorial assistance.

References

Acosta-Serrano, A., Almeida, I.C., Freitas-Junior, L.H., Yoshida, N., Schenkman, S., 2001. The mucin-like glycoprotein super-family of Trypanosoma cruzi: structure and biological roles. Mol Biochem Parasitol 114, 143–150.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402.

Andersen, J.F., Francischetti, I.M., Valenzuela, J.G., Schuck, P., Ribeiro, J.M., 2003. Inhibition of hemostasis by a high-affinity biogenic amine-binding protein from the saliva of a blood-feeding insect. J Biol Chem 278, 4611–4617.

Arca, B., Lombardo, F., de Lara Capurro, M., della Torre, A., Dimopoulos, G., James, A.A., Coluzzi, M., 1999. Trapping cDNAs encoding secreted proteins from the salivary glands of the malaria vector Anopheles gambiae. Proc Natl Acad Sci USA 96, 1516–1521.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25–29.

Ashburner, M., Hoy, M.A., Peloquin, J.J., 1998. Prospects for the genetic transformation of arthropods. Insect Mol Biol 7, 201–213.

Barral, A., Honda, E., Caldas, A., Costa, J., Vinhas, V., Rowton, E.D., Valenzuela, J.G., Charlab, R., Barral-Netto, M., Ribeiro, J.M., 2000. Human immune response to sand fly salivary gland antigens: a useful epidemiological marker? Am J Trop Med Hyg 62, 740–745.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., Sonnhammer, E.L., 2000. The Pfam protein families database. Nucleic Acids Res 28, 263–266.

Catteruccia, F., Nolan, T., Loukeris, T.G., Blass, C., Savakis, C., Kafatos, F.C., Crisanti, A., 2000. Stable germline transformation of the malaria mosquito Anopheles stephensi. Nature 405, 959–962.

Champagne, D.E., Smartt, C.T., Ribeiro, J.M., James, A.A., 1995. The salivary gland-specific apyrase of the mosquito Aedes aegypti is a member of the 5′-nucleotidase family. Proc Natl Acad Sci USA 92, 694–698.

Charlab, R., Valenzuela, J.G., Rowton, E.D., Ribeiro, J.M., 1999.

Toward an understanding of the biochemical and pharmacological complexity of the saliva of a hematophagous sand fly Lutzomyia longipalpis. Proc Natl Acad Sci USA 96, 15155–15160.

Dimopoulos, G., Richman, A., Muller, H.M., Kafatos, F.C., 1997. Molecular immune responses of the mosquito Anopheles gambiae to bacteria and malaria parasites. Proc Natl Acad Sci USA 94, 11508–11513.

Francischetti, I.M., Andersen, J.F., Ribeiro, J.M., 2002a. Biochemical and functional characterization of recombinant Rhodnius prolixus platelet aggregation inhibitor 1 as a novel lipocalin with high affinity for adenosine diphosphate and other adenine nucleotides. Biochemistry 41, 3810–3818.

Francischetti, I.M., Valenzuela, J.G., Andersen, J.F., Mather, T.N., Ribeiro, J.M., 2002b. Ixolaris, a novel recombinant tissue factor pathway inhibitor (TFPI) from the salivary gland of the tick, Ixodes scapularis: identification of factor X and factor Xa as scaffolds for the inhibition of factor VIIa/tissue factor complex. Blood 99, 3602–3612.

Francischetti, I.M., Valenzuela, J.G., Pham, V.M., Garfield, M.K., Ribeiro, J.M., 2002c. Toward a catalog for the transcripts and proteins (sialome) from the salivary gland of the malaria vector Anopheles gambiae. J Exp Biol 205, 2429–2451.

Francischetti, I.M., Valenzuela, J.G., Ribeiro, J.M., 1999. Anophelin: kinetics and mechanism of thrombin inhibition. Biochemistry 38, 16678–16685.

Gompertz, S., Stockley, R.A., 2000. Inflammation--role of the neutrophil and the eosinophil. Semin Respir Infect 15, 14–23.

Gooding, R.H., 1972. Digestive process of haematophagous insects I-A literature review. Quaest Ent 8, 5–60.

Grossman, G.L., Rafferty, C.S., Clayton, J.R., Stevens, T.K., Mukabayire, O., Benedict, M.Q., 2001. Germline transformation of the malaria vector, Anopheles gambiae, with the piggyBac transposable element. Insect Mol Biol 10, 597–604.

Hansen, J.E., Lund, O., Tolstrup, N., Gooley, A.A., Williams, K.L., Brunak, S., 1998. NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. Glycoconj J 15, 115–130.

Hati, A.K., 1997. Urban malaria vector biology. Indian J Med Res 106, 149–163.

Hekmat-Scafe, D.S., Dorit, R.L., Carlson, J.R., 2000. Molecular evolution of odorant-binding protein genes OS-E and OS-F in Drosophila. Genetics 155, 117–127.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., et al. 2002. The genome sequence of the malaria mosquito Anopheles gambiae. Science 298, 129–149.

Huang, X., 1992. A contig assembly program based on sensitive detection of fragment overlaps. Genomics 14, 18–25.

Hvidsten, T.R., Komorowski, J., Sandvik, A.K., Laegreid, A., 2001. Predicting gene function from gene expressions and ontologies. Pac Symp Biocomput, 299–310.

Isawa, H., Yuda, M., Orito, Y., Chinzei, Y., 2002. A mosquito salivary protein inhibits activation of the plasma contact system by binding to factor XII and high molecular weight kininogen. J Biol Chem 13, 13.

James, A.A., Blackmer, K., Marinotti, O., Ghosn, C.R., Racioppi, J.V., 1991. Isolation and characterization of the gene expressing the major salivary gland protein of the female mosquito, Aedes aegypti. Mol Biochem Parasitol 44, 245–253.

James, A.A., Blackmer, K., Racioppi, J.V., 1989. A salivary gland-specific, maltase-like gene of the vector mosquito, Aedes aegypti. Gene 75, 73–83.

Lanfrancotti, A., Lombardo, F., Santolamazza, F., Veneri, M., Castrignano, T., Coluzzi, M., Arca, B., 2002. Novel cDNAs encoding salivary proteins from the malaria vector Anopheles gambiae. FEBS Lett 517, 67–71.

Lanzaro, G.C., Lopes, A.H., Ribeiro, J.M., Shoemaker, C.B., Warburg, A., Soares, M., Titus, R.G., 1999. Variation in the salivary peptide, maxadilan, from species in the Lutzomyia longipalpis complex. Insect Mol Biol 8, 267–275.

Lewis, S., Ashburner, M., Reese, M.G., 2000. Annotating eukaryote genomes. Curr Opin Struct Biol 10, 349–354.

Li, S., Kwon, J., Aksoy, S., 2001. Characterization of genes expressed in the salivary glands of the tsetse fly, Glossina morsitans morsitans. Insect Mol Biol 10, 69–76.

Lombardo, F., Di Cristina, M., Spanos, L., Louis, C., Coluzzi, M., Arca, B., 2000. Promoter sequences of the putative Anopheles gambiae apyrase confer salivary gland expression in Drosophila melanogaster. J Biol Chem 275, 23861–23868.

Marinotti, O., James, A., Ribeiro, J.M.C., 1990. Diet and salivation in female Aedes aegypti mosquitoes. J Insect Physiol 36, 545–548.

Mathews, G.V., Sidjanski, S., Vanderberg, J.P., 1996. Inhibition of mosquito salivary gland apyrase activity by antibodies produced in mice immunized by bites of Anopheles stephensi mosquitoes. Am J Trop Med Hyg 55, 417–423.

Metcalf, R.L., 1945. The physiology of the salivary glands of Anopheles quadrimaculatus. J Nat Malaria Soc 4, 271.

Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10, 1–6.

Nolan, T., Bower, T.M., Brown, A.E., Crisanti, A., Catteruccia, F., 2002. piggyBac-mediated germline transformation of the malaria mosquito Anopheles stephensi using the red fluorescent protein dsRED as a selectable marker. J Biol Chem 277, 8759–8762.

Paesen, G.C., Adams, P.L., Nuttall, P.A., Stuart, D.L., 2000. Tick histamine-binding proteins: lipocalins with a second binding cavity. Biochim Biophys Acta 1482, 92–101.

Ribeiro, J.M., Francischetti, I.M., 2001. Platelet-activating-factor-hydrolyzing phospholipase C in the salivary glands and saliva of the mosquito Culex quinquefasciatus. J Exp Biol 204, 3887–3894.

Ribeiro, J.M., Valenzuela, J.G., 1999. Purification and cloning of the salivary peroxidase/catechol oxidase of the mosquito Anopheles albimanus. J Exp Biol 202, 809–816.

Ribeiro, J.M.C., 1995. Blood-feeding arthropods: Live syringes or invertebrate pharmacologists? Infect Agents Dis 4, 143–152.

Ribeiro, J.M.C., Nussenzveig, R.H., 1993. The salivary catechol oxidase/peroxidase activities of the mosquito, Anopheles albimanus. J Exp Biol 179, 273–287.

Ribeiro, J.M.C., Nussenzveig, R.H., Tortorella, G., 1994. Salivary vasodilators of Aedes triseriatus and Anopheles gambiae (Diptera: Culicidae). J Med Entomol 31, 747–753.

Ropert, C., Ferreira, L.R., Campos, M.A., Procopio, D.O., Travassos, L.R., Ferguson, M.A., Reis, L.F., Teixeira, M.M., Almeida, I.C., Gazzinelli, R.T., 2002. Macrophage signaling by glycosylphosphatidylinositol-anchored mucin-like glycoproteins derived from Trypanosoma cruzi trypomastigotes. Microbes Infect 4, 1015–1025.

Rossignol, P.A., Lueders, A.M., 1986. Bacteriolytic factor in the salivary glands of Aedes aegypti. Comp Biochem Physiol 83B, 819–822.

Sankoff, D., 2001. Gene and genome duplication. Curr Opin Genet Dev 11, 681–684.

Schreiber, M.C., Karlo, J.C., Kovalick, G.E., 1997. A novel cDNA from Drosophila encoding a protein with similarity to mammalian cysteine-rich secretory proteins, wasp venom antigen 5, and plant group 1 pathogenesis-related proteins. Gene 191, 135–141.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., Bork, P., 2000. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res 28, 231–234.

Schwartz, B.S., Ribeiro, J.M.C., Goldstein, M.D., 1990. Anti-tick antibodies: An epidemiological tool in Lyme disease research. Am J Epidemiol 132, 58–66.

Simons, F.E., Peng, Z., 2001. Mosquito allergy: recombinant mosquito salivary antigens for new diagnostic tests. Int Arch Allergy Immunol 124, 403–405.

Soares, R.P., Gontijo, N.F., Romanha, A.J., Diotaiuti, L., Pereira, M.H., 1998. Salivary heme proteins distinguish Rhodnius prolixus from Rhodnius robustus (Hemiptera: Reduviidae: Triatominae). Acta Trop 71, 285–291.

Soares, R.P., Sant'Anna, M.R., Gontijo, N.F., Romanha, A.J., Diotaiuti, L., Pereira, M.H., 2000. Identification of morphologically similar Rhodnius species (Hemiptera: Reduviidae: Triatominae) by electrophoresis of salivary heme proteins. Am J Trop Med Hyg 62, 157–161.

Valenzuela, J.G., Charlab, R., Gonzalez, E.C., Miranda-Santos, I.K.F., Marinotti, O., Francischetti, I.M., Ribeiro, J.M.C., 2002a. The D7 family of salivary proteins in blood sucking Diptera. Insect Mol Biol 11, 149–155.

Valenzuela, J.G., Francischetti, I.M., Ribeiro, J.M., 1999. Purification, cloning, and synthesis of a novel salivary anti-thrombin from the mosquito Anopheles albimanus. Biochemistry 38, 11209–11215.

Valenzuela, J.G., Francischetti, I.M.B., Pham, V.M., Garfield, M.K., Mather, T.N., Ribeiro, J.M.C., 2002b. Exploring the sialome of the tick, Ixodes scapularis. J Exp Biol 205, 2843–2864.

Valenzuela, J.G., Pham, V.M., Garfield, M.K., Francischetti, I.M., Ribeiro, J.M.C., 2002c. Toward a description of the sialome of the adult female mosquito Aedes aegypti. Insect Biochem Mol Biol 32, 1101–1122.

Vasta, G.R., Quesenberry, M., Ahmed, H., O'Leary, N., 1999. C-type lectins and galectins mediate innate and adaptive immune functions: their roles in the complement activation pathway. Dev Comp Immunol 23, 401–420.

Vaughan, J.A., Noden, B.H., Beier, J.C., 1991. Concentration of human erythrocytes by anopheline mosquitoes (Diptera:Culicidae) during feeding. J Med Entomol 28, 780–786.

Zdobnov, E.M., Von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R.R., Christophides, G.K., Thomasova, D., Holt, R.A., Subramanian, G.M., et al. 2002. Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. Science 298, 149–159.